Partition models and cluster processes

With applications to classification

Peter McCullagh Jie Yang

Department of Statistics
University of Chicago

TTI, Chicago, November 6 2006

Outline

Fisher Discriminant model

Description as a process (i.i.d.)

k classes with frequencies $\pi_1, \ldots, \pi_k$
Features X for class r are iid $N(\mu_r, \Sigma)$ on $\mathcal{X} = \mathcal{R}^q$

Observations $(Y_1, X_1), (Y_2, X_2), \ldots$ are i.i.d
$\mathrm{pr}(Y_i = r) = \pi_r$ i.i.d.
Given Y, features are iid $X_j \sim N(\mu_{y_j}, \Sigma)$

Density of (Y_i, X_i) at (y, x)

$$\pi_y |\Sigma|^{-q/2} \exp\big(-(x - \mu_y)' \Sigma^{-1} (x - \mu_y)/2\big)$$

Fisher Discriminant model

Application to classification

Estimation step:

$$\hat{\pi}_r = n_r/n, \quad \hat{\mu}_r = \bar{x}_r, \quad \hat{\Sigma} = S;$$

Classification step:

for a new unit u' such that $X(u') = x'$

$\mathrm{pr}(Y(u') = r \,|\, \text{data}) \propto \pi_r \phi(x' - \mu_r; \Sigma)$

(independent of data other than x')

Substitution step:

$\hat{\mathrm{pr}}(Y(u') = r \,|\, \text{data}) \propto n_r \phi(x' - \bar{x}_r; S)$

(dependent on data through estimate)

Logistic regression model

Description as a process (indep but not i.d.)

Feature $x(u)$ regarded as a covariate for unit u

Class $Y(u)$ regarded as 'response'

Linear logistic model:

Components $Y(u_1), Y(u_2), \ldots$ independent

$\log \mathrm{pr}(Y(u) = r) = \alpha_r + \beta_r x(u) + \text{const}$
consistent with Fisher model $\beta_r = \Sigma^{-1}\mu_r$

Logistic regression model

Application to classification

Estimation step:
maximum likelihood gives $\hat{\alpha}, \hat{\beta}$

Classification step:

$$\mathrm{pr}(Y(u') = r \,|\, \text{data}) = \frac{\exp(\alpha_r + \beta_r x(u'))}{\sum_c \exp(\alpha_c + \beta_c x(u'))}$$

(independent of data other than $x(u')$)

Substitution step: $\beta \mapsto \hat{\beta}$ (fitted = predicted)

Random partition of $[n] = \{1, \ldots, n\}$

A set of disjoint non-empty subsets whose union is $[n]$
$n = 2$: $B = \{\{1, 2\}\} = 12$ or $B = \{\{1\}, \{2\}\} = 1|2$
$n = 3$: $\mathcal{B}_3 = \{123, 1|23, 2|13, 3|12, 1|2|3\}$
$n = 4$: $\mathcal{B}_4 = \{1234, 123|4[4], 12|34[3], 12|3|4[6], 1|2|3|4\}$.
Bell numbers: $\#\mathcal{B}_3 = 5$; $\#\mathcal{B}_4 = 15$; $\#\mathcal{B}_5 = 52$; $\#B_6 = 203$

$$B = 12|34|5 \equiv \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathcal{B}_5$$

Permutation of elements: $B^\sigma_{ij} = B_{\sigma(i)\sigma(j)}$

$$B = 12|34|5 \equiv \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathcal{B}_5$$

Example 1: $\sigma = (1,3)$: $\quad B^\sigma = 32|14|5$

Example 2: $\sigma = (2,4,5)$: $\quad B^\sigma = 15|23|4$

$B^\sigma(1,5) = B(1,2) = 1, \quad B^\sigma(1,2) = B(1,4) = 0$

Block sizes unaffected

Exchangeable product partition distribution (Hartigan 1990)

$$p_n(B) = D_n \prod_{b \in B} C(\#b)$$

Deletion of elements

Deletion (of 5) from $12|34|5$ gives $12|34$ in $\mathcal{B}_4$
Deletion (of 5) from $135|2|4$ gives $13|2|4$ in $\mathcal{B}_4$

Deletion is a map $D : \mathcal{B}_n \to \mathcal{B}_{n-1}$
How does this affect distributions?

Example: p_5 is uniform on the 15 elements of $\mathcal{B}_5$
Marginal distribution on $\mathcal{B}_4$ after deletion is
$p_4(1234) = p_5(12345) + p_5(1234|5) = 2/15$
$p_4(123|4) = p_5(1235|4) + p_5(123|45) + p_5(123|4|5) = 3/15$
$p_4(12|3|4) = p_5(125|3|4) + \cdots + p_5(12|3|4|5) = 4/15$
Marginal distribution not uniform.

Ewens partition process

$$p_n(B) = \frac{\lambda^{\#B}\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{b \in B} \Gamma(\#b)$$

Exponential family of distns on $\mathcal{B}_n$
- canonical parameter $\theta = \log \lambda$
- canonical statistic $\#B$: $\quad E(\#B) \simeq \lambda \log(n)$
- cumulant function $\log \Gamma(n + e^\theta) - \log \Gamma(e^\theta)$

Product partition model: $C(\#b) = \lambda \times (\#b - 1)!$.
Finitely exchangeable for each n
p_n is marginal distn of p_{n+1}

distribution for $n = 4$

$$
\begin{aligned}
p_4(1234) &= 6/(\lambda+1)(\lambda+2)(\lambda+3), \\
p_4(123|4) &= 2\lambda/(\lambda+1)(\lambda+2)(\lambda+3), \\
p_4(12|34) &= \lambda/(\lambda+1)(\lambda+2)(\lambda+3) \\
p_4(12|3|4) &= \lambda^2/(\lambda+1)(\lambda+2)(\lambda+3), \\
p_4(1|2|3|4) &= \lambda^3/(\lambda+1)(\lambda+2)(\lambda+3)
\end{aligned}
$$

$$
\begin{aligned}
p_3(123) &= p_4(1234) + p_4(123|4) = 2/(\lambda+1)(\lambda+2) \\
p_3(12|3) &= p_4(124|3) + p_4(12|34) + p_4(12|3|4) \\
p_3(1|2|3) &= p_4(14|2|3) + p_4(1|24|3) + p_4(1|2|34) + p_4(1|2|3|4)
\end{aligned}
$$

p_n is marginal distribution of p_{n+1}

Infinitely exchangeable partition process

Chinese restaurant process

Ewens process: $B \equiv B_\infty$ is a random partition of natural numbers Regard as a sequence of matrices $B_1, B_2, \ldots$ such that $B_n(i,j) = B_\infty(i,j)$ is leading sub-matrix of order n.
$B_n \sim \text{Ewens}(n, \lambda)$
Conditional distribution of B_{n+1} given B_n
Transition distribution $B_n \to B_{n+1}$

$$\text{pr}(u_{n+1} \mapsto b | B_n) = \begin{cases} \#b/(n+\lambda) & b \in B_n \\ \lambda/(n+\lambda) & b = \emptyset \end{cases}$$

CRP interpretation due to Dubins and Pitman
No of occupied tables is approx Poisson($\lambda \log(n)$)

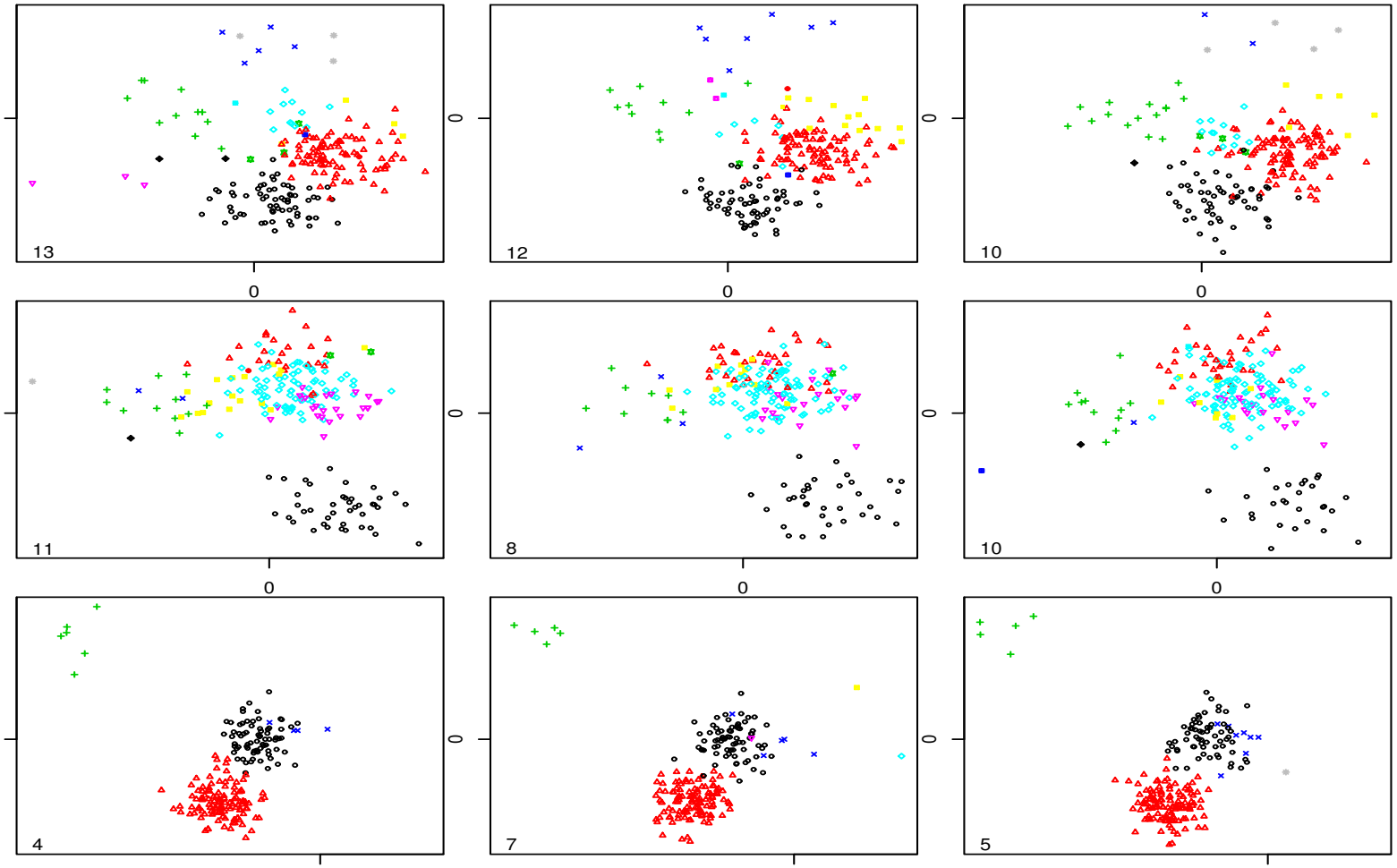
Gauss Ewens cluster processes

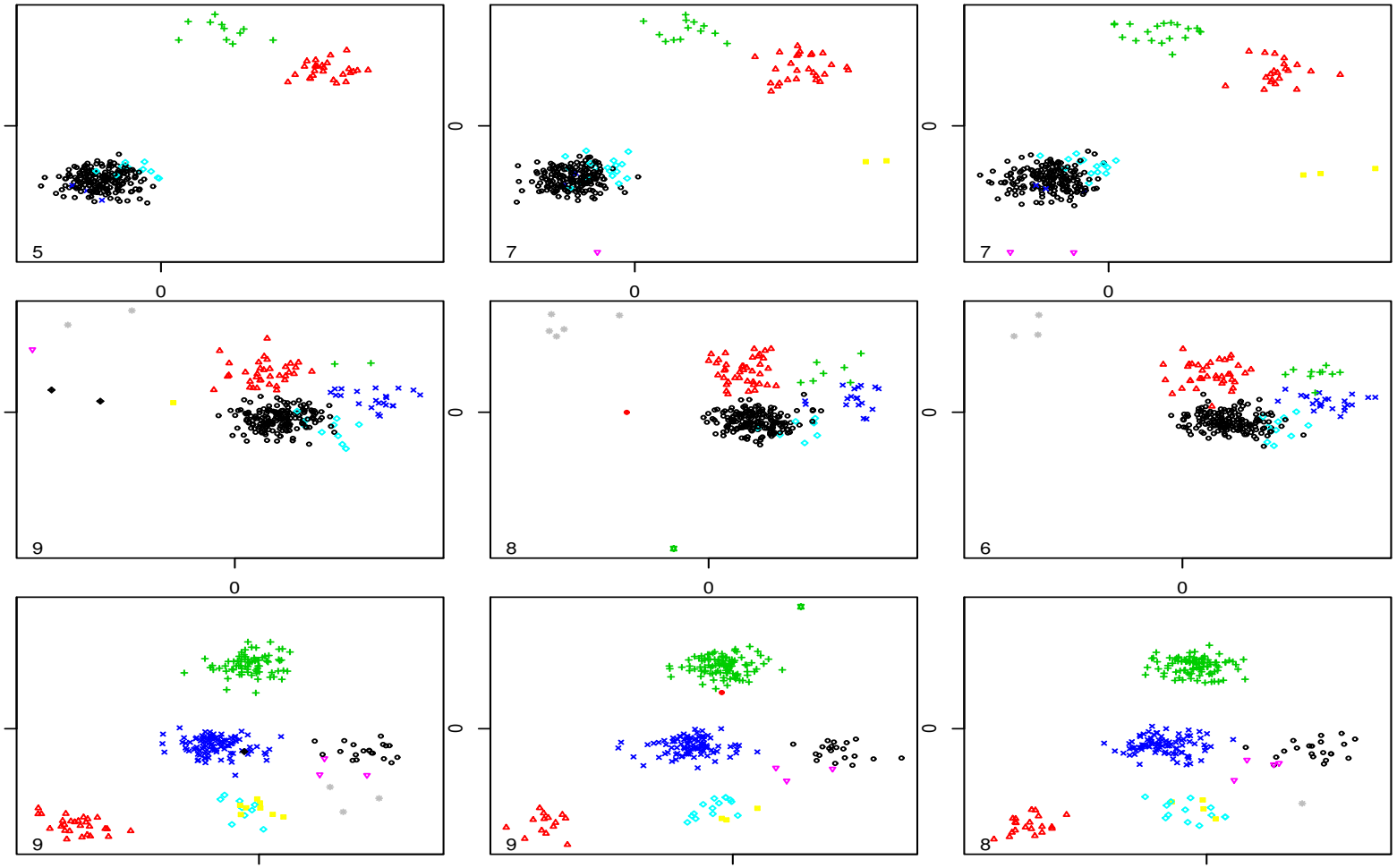
Pair (B, X) such that $X = (X_1, X_2, \ldots)$ (random sequence)
$B = \{B_{ij}\}$ is a random partition of integers

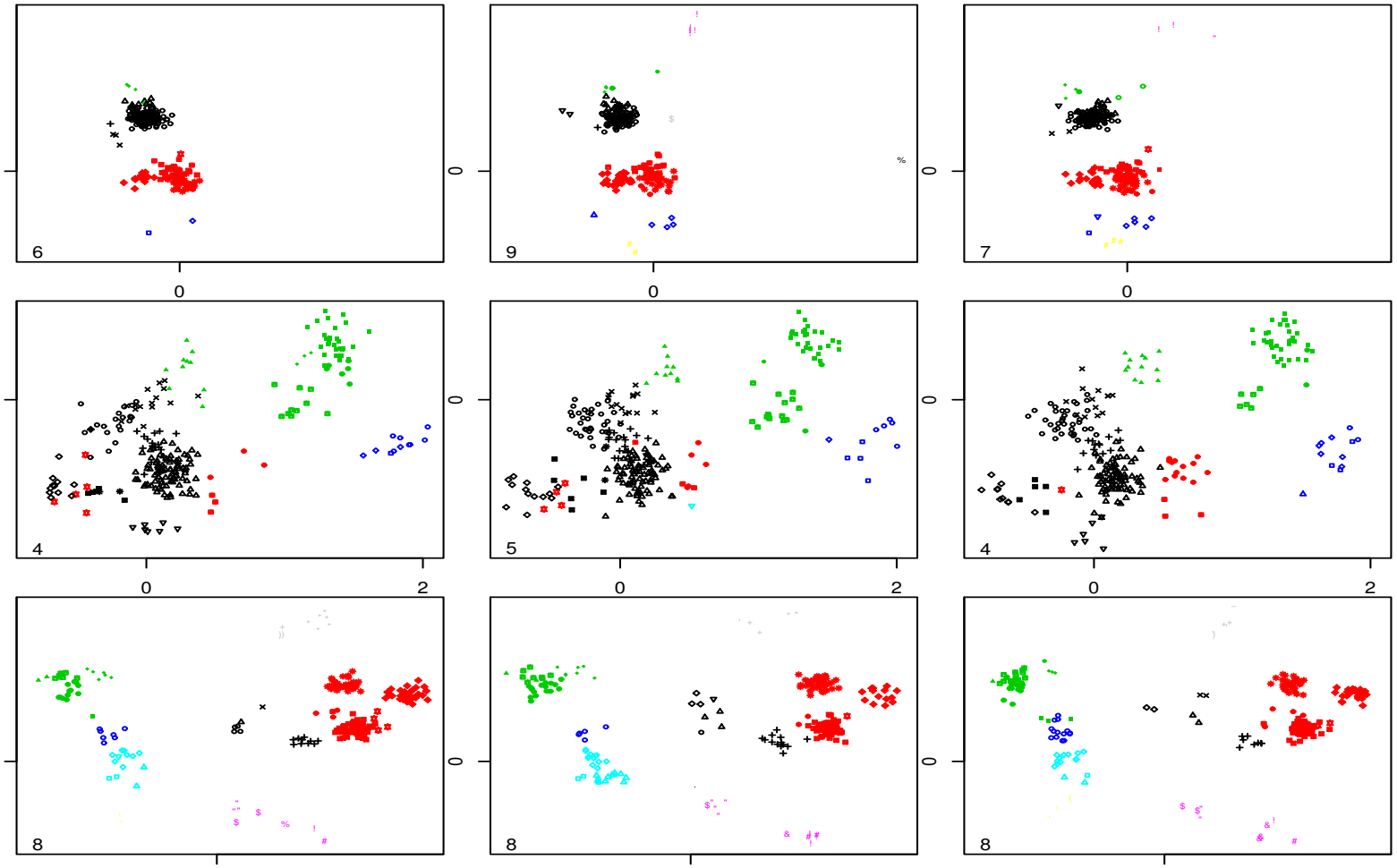
Finite-dimensional distributions
$p_n(B, x)$ is distn on $\mathcal{B}_n \times \mathcal{X}^n$
Invariant under permutation $p_n(B^\sigma, x^\sigma) = p_n(B, x)$
p_n is the marginal of p_{n+1}

$B \sim \text{Ewens}_n(\lambda)$; given B, $X \sim N(0, \Sigma = I_n + \theta B)$

$$p_n(B, x) = \frac{\lambda^{\#B}\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{b \in B} \Gamma(\#b) \times |2\pi\Sigma|^{-1/2} \exp(-x'\Sigma^{-1}x/2)$$


13
12
10
11
8
10
4
7
5


5
7
7
9
8
6
9
9
8


6
9
7
4
5
4
8
8
8
0
2

Potential applications of cluster process

- Prediction: given $(B, X)^{(n)}$ predict next value
  — Easily computed but not interesting or useful
- Density estimation: Given $X_1, \ldots, X_n$ (No B)
  predict next value — $p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)$
- Cluster analysis: $p_n(B|X)$ is a distn on clusterings of [n]
  $p_n(B_{i,j} = 1|X) = E(B_{i,j}|X)$ (triples too)
  $p_n(\#B = r|X)$ (conditional distn of $\#B$)
  — preliminary conclusions ...
- Classification: given $(B, X)^{(n)}$ and $X(u_{n+1}) = x'$
  compute $p_{n+1}(u_{n+1} \mapsto b|\text{data})$
  — Easily computed and fairly useful

Potential applications of cluster process

- Prediction: given $(B, X)^{(n)}$ predict next value
  — Easily computed but not interesting or useful
- Density estimation: Given $X_1, \ldots, X_n$ (No B)
  predict next value — $p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)$
- Cluster analysis: $p_n(B|X)$ is a distn on clusterings of [n]
  $p_n(B_{i,j} = 1|X) = E(B_{i,j}|X)$ (triples too)
  $p_n(\#B = r|X)$ (conditional distn of $\#B$)
  — preliminary conclusions ...
- Classification: given $(B, X)^{(n)}$ and $X(u_{n+1}) = x'$
  compute $p_{n+1}(u_{n+1} \mapsto b|\text{data})$
  — Easily computed and fairly useful

Potential applications of cluster process

- Prediction: given $(B, X)^{(n)}$ predict next value
  — Easily computed but not interesting or useful
- Density estimation: Given $X_1, \ldots, X_n$ (No B)
  predict next value — $p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)$
- Cluster analysis: $p_n(B|X)$ is a distn on clusterings of [n]
  $p_n(B_{i,j} = 1|X) = E(B_{i,j}|X)$ (triples too)
  $p_n(\#B = r|X)$ (conditional distn of $\#B$)
  — preliminary conclusions ...
- Classification: given $(B, X)^{(n)}$ and $X(u_{n+1}) = x'$
  compute $p_{n+1}(u_{n+1} \mapsto b|\text{data})$
  — Easily computed and fairly useful

Potential applications of cluster process

- Prediction: given $(B, X)^{(n)}$ predict next value
  — Easily computed but not interesting or useful
- Density estimation: Given $X_1, \ldots, X_n$ (No B)
  predict next value — $p_{n+1}(X_{n+1} \mid X_1, \ldots, X_n)$
- Cluster analysis: $p_n(B|X)$ is a distn on clusterings of [n]
  $p_n(B_{i,j} = 1|X) = E(B_{i,j}|X)$ (triples too)
  $p_n(\#B = r|X)$ (conditional distn of $\#B$)
  — preliminary conclusions ...
- Classification: given $(B, X)^{(n)}$ and $X(u_{n+1}) = x'$
  compute $p_{n+1}(u_{n+1} \mapsto b|\text{data})$
  — Easily computed and fairly useful

Classification and cluster processes

Given $(B, \mathbf{x})$ on a sample of n units and
a new unit u' with $x(u') = x'$.
Classify the new unit (conditional distribution)

Gauss-Ewens conditional distribution

$$\mathrm{pr}(u' \mapsto b \,|\, \ldots) \propto \begin{cases} (\#b)\, \phi_{n+1}(\mathbf{x}'; I_{n+1} + \theta B_b) & b \in B \\ \lambda\, \phi_{n+1}(\mathbf{x}'; I_{n+1} + \theta B_\emptyset) & b = \emptyset \end{cases}$$

$\mathbf{x}' = (\mathbf{x}, x')$; $B_b = \{B, u' \mapsto b\}$

Application to classification (contd)

Simplification of conditional distribution

$$\mathrm{pr}(u' \mapsto b \,|\, \ldots) \propto \begin{cases} (\#b)\, \phi_1(x(u') - \tilde{\mu}_b; \tilde{\sigma}_b^2) & b \in B \\ \lambda\, \phi_1(x(u'); 1+\theta) & b = \emptyset \end{cases}$$

$$\tilde{\mu}_b = \bar{x}_b n_b \theta/(1+n_b\theta), \quad \tilde{\sigma}_b^2 = 1 + \theta/(1+n_b\theta)$$

Typical values $\theta \geq 5$ and $n_b \geq 5$
so $\tilde{\mu}_b/\bar{x}_b = n_b\theta/(1+n_b\theta) \geq 0.96$

(similar to Fisher discriminant model, but with shrinkage)

Cox process models

Construction:

$\lambda(x)$ is a random intensity function on $\mathcal{X} = \mathcal{R}$

$\mathbf{X}$ is a Poisson process with intensity λ.

Product density:

$m(x) = E(\lambda(x))$ first-order p.d.

$m(x_1, x_2) = E\big(\lambda(x_1)\lambda(x_2)\big)$ second order p.d.

$m(\mathbf{x}) = E\big(\lambda(x_1) \cdots \lambda(x_n)\big)$ is nth order p.d.

$\mathbf{X}(A) = \#(X \cap A)$ = no of events in A

$m(\mathbf{x})\, d\mathbf{x}$: expected no of ordered n-tuples of events in $d\mathbf{x}$

Marked Cox process

$\lambda_1(x), \ldots, \lambda_k(x)$ indep random intensity functions
$\mathbf{X}^{(r)}$ Cox process with intensity $\lambda_r(x)$
$\mathbf{X} = \cup \mathbf{X}^{(r)}$ superposition process
Marked process is labelled partition $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(k)})$ of $\mathbf{X}$

$\mathrm{pr}(\mathbf{X}^{(r)}$ contains event in $d\mathbf{x}^{(r)}) = m_r(d\mathbf{x}^{(r)}) + o(d\mathbf{x}^{(r)})$

$\mathrm{pr}(\mathbf{X}$ contains event in $d\mathbf{x}) = m(d\mathbf{x}) + o(d\mathbf{x})$

$$\mathrm{pr}(y \mid \mathbf{x} \subset \mathbf{X}) = \frac{\prod m_r(\mathbf{x}^{(r)})}{m(\mathbf{x})}$$

label $y \colon \mathbf{x} \to \mathcal{C}$ and $\mathbf{x}^{(r)} = y^{-1}(r)$

Boson and related processes

$Z(x)$ is a GP $N(0, K)$ on $\mathcal{X}$ (complex-valued for simplicity!)
$\lambda(x) = |Z(x)|^2$ is intensity at x for Cox process

product density: $m(\mathbf{x}) = E(\lambda(x_1) \cdots \lambda(x_n)) = \mathrm{per}[K](\mathbf{x})$
matrix $[K](\mathbf{x}) = \{K(x_i, x_j)\}$

Permanent polynomial of matrix A:

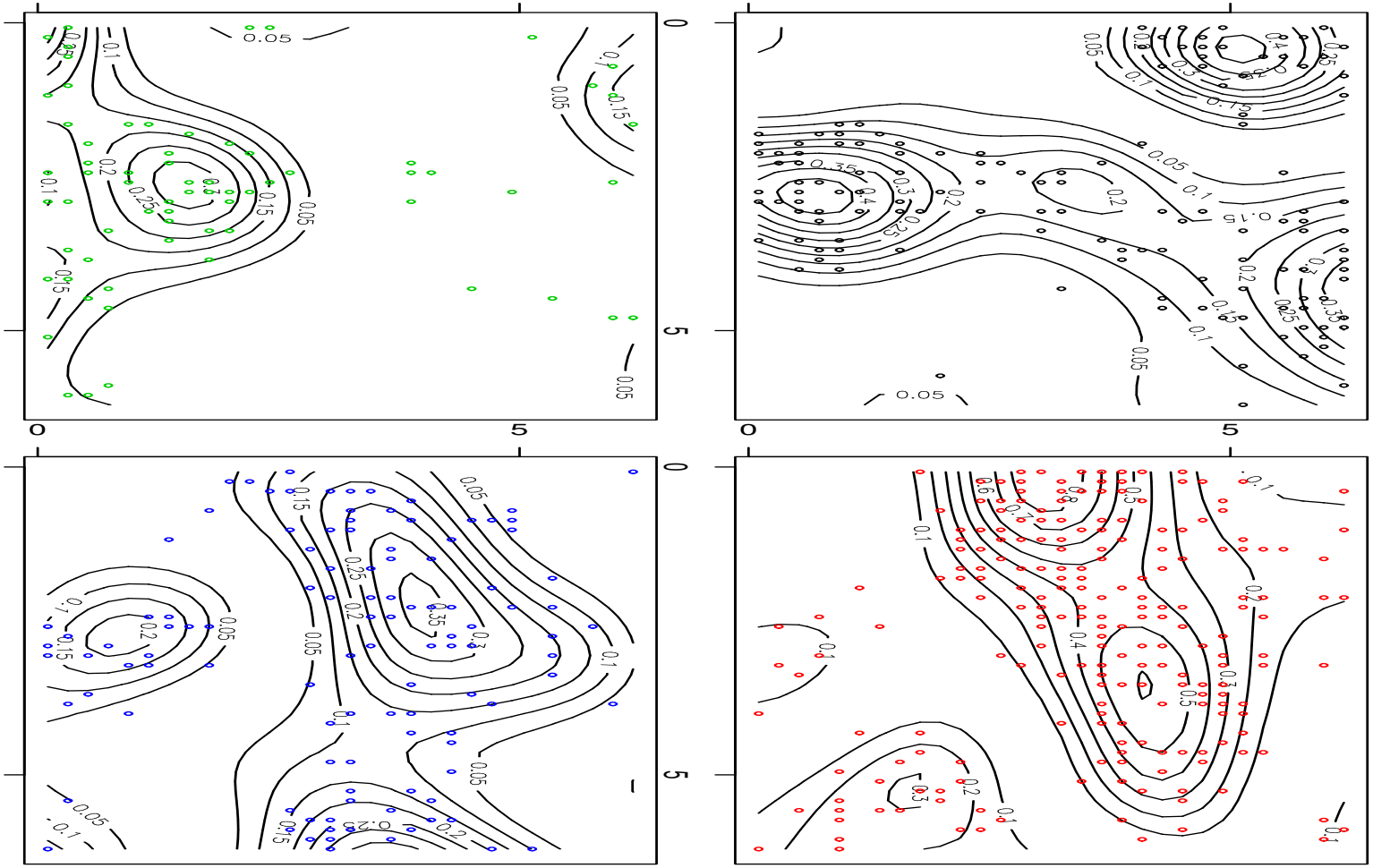
$$\mathrm{per}_\alpha(A) = \sum_\sigma \alpha^{\#\sigma} A_{1,\sigma(1)} \cdots A_{n,\sigma(n)}$$

Convolution or superposition:

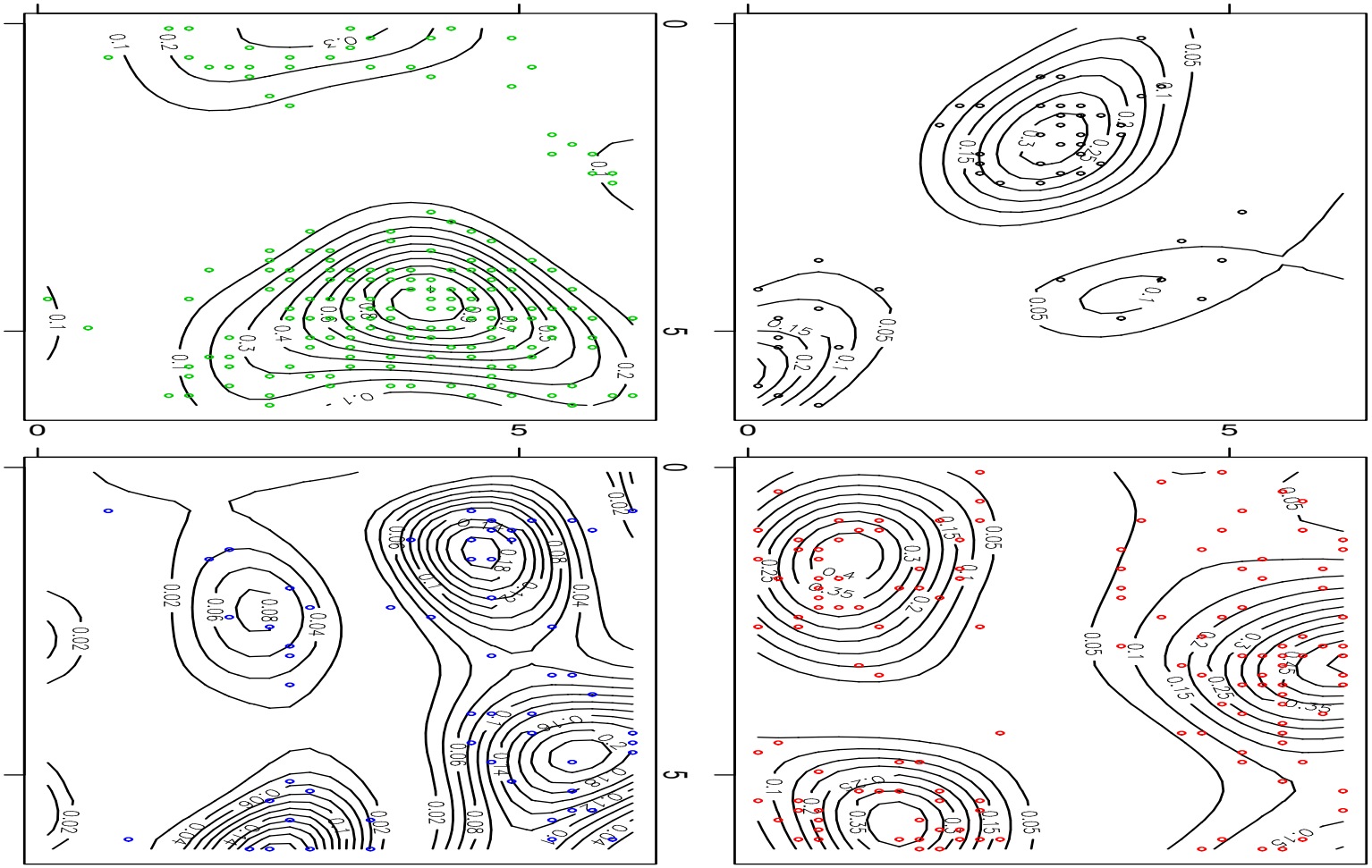
$$\sum_{\mathbf{w} \subset \mathbf{x}} \mathrm{per}_\alpha[K](\mathbf{w}) \, \mathrm{per}_{\alpha'}[K](\bar{\mathbf{w}}) = \mathrm{per}_{\alpha+\alpha'}[K](\mathbf{x})$$

Macchi 1975; Shirai and Takahashi 2003; McCullagh and Møller 2006

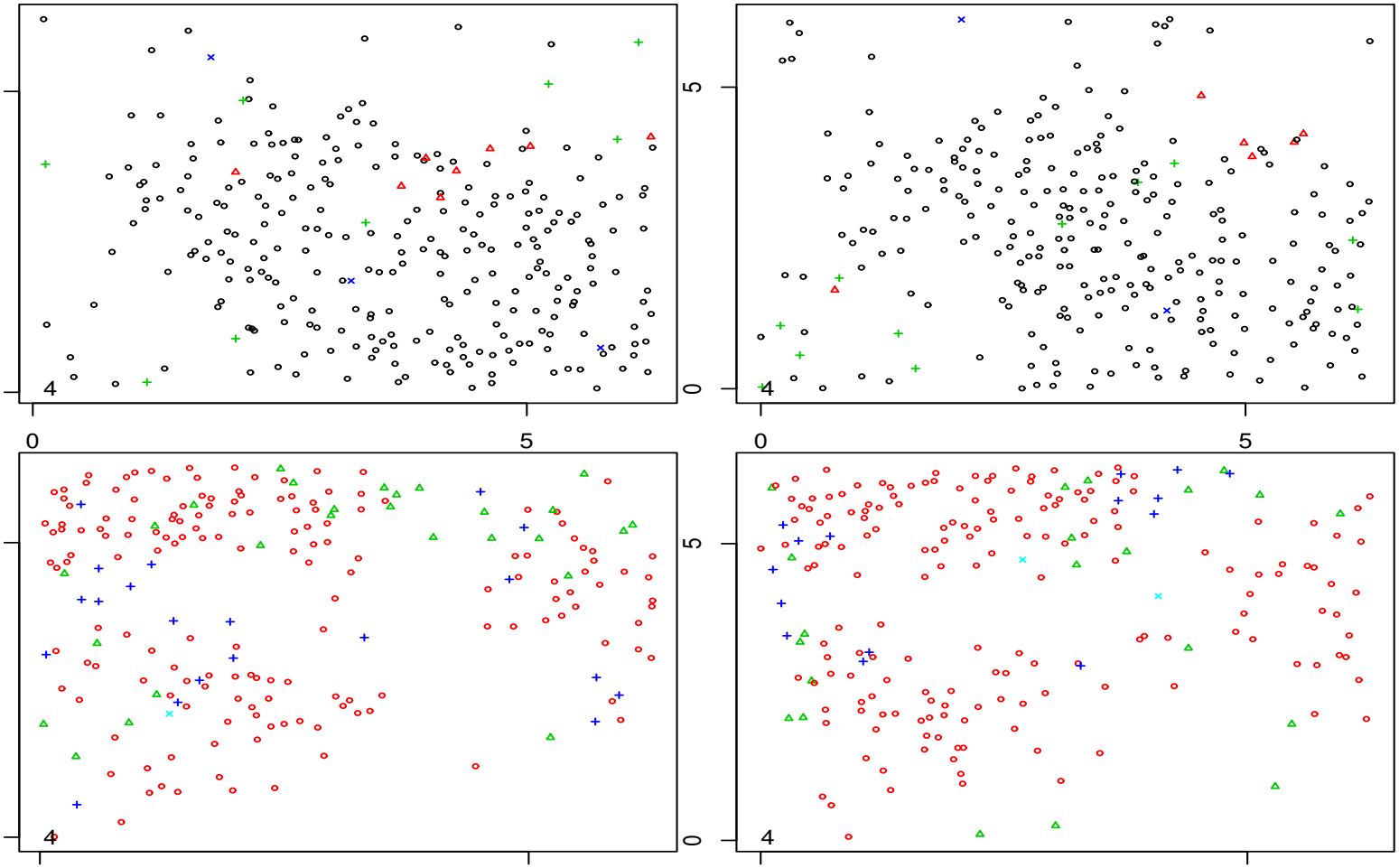
Intensity functions and Cox processes



Intensity functions and Cox processes



Superposition processes



Permanent partition models

Given that $\mathbf{x} \subset \mathbf{X}$ occurs in the superposition
Conditional distribution of labels $y \colon \mathbf{x} \to \mathcal{C}$

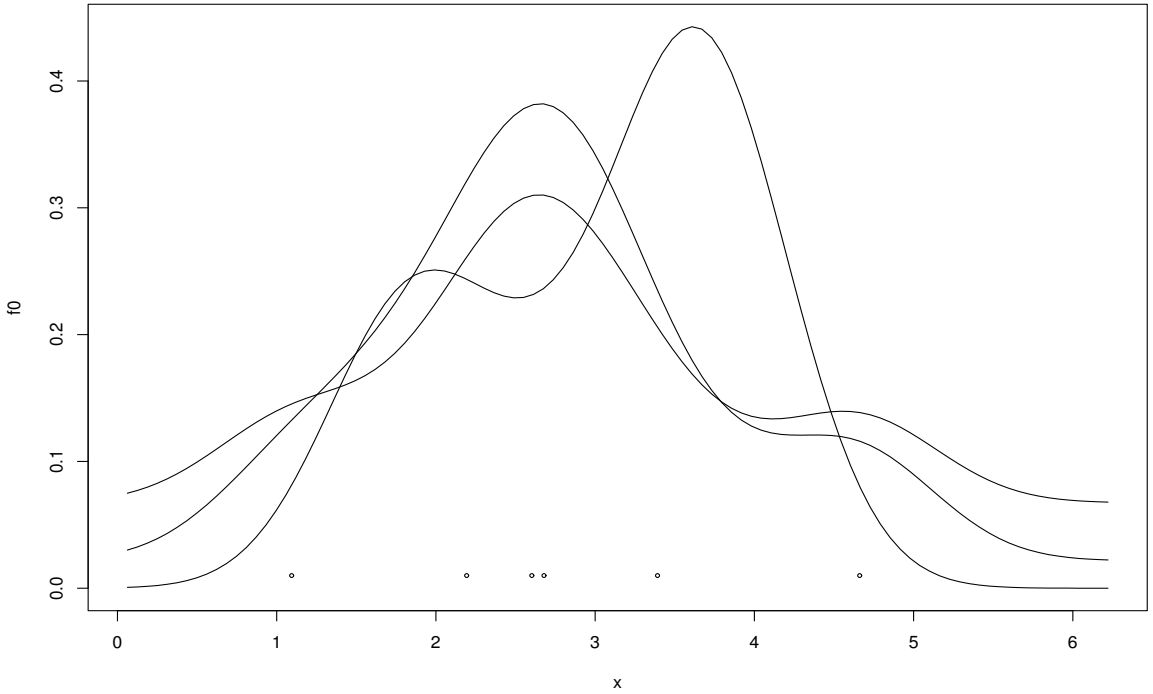
$$p_n(y \mid \mathbf{x}) = \frac{\mathrm{per}_{\alpha_1}[K](\mathbf{x}^{(1)}) \cdots \mathrm{per}_{\alpha_k}[K](\mathbf{x}^{(k)})}{\mathrm{per}_{\alpha_\bullet}[K](\mathbf{x})}$$

is a prob distribution on labelled partitions of $\mathbf{x}$.
reduces to Dirichlet multinomial if $K(x, x') \equiv 1$

Classification distribution for new event at x'

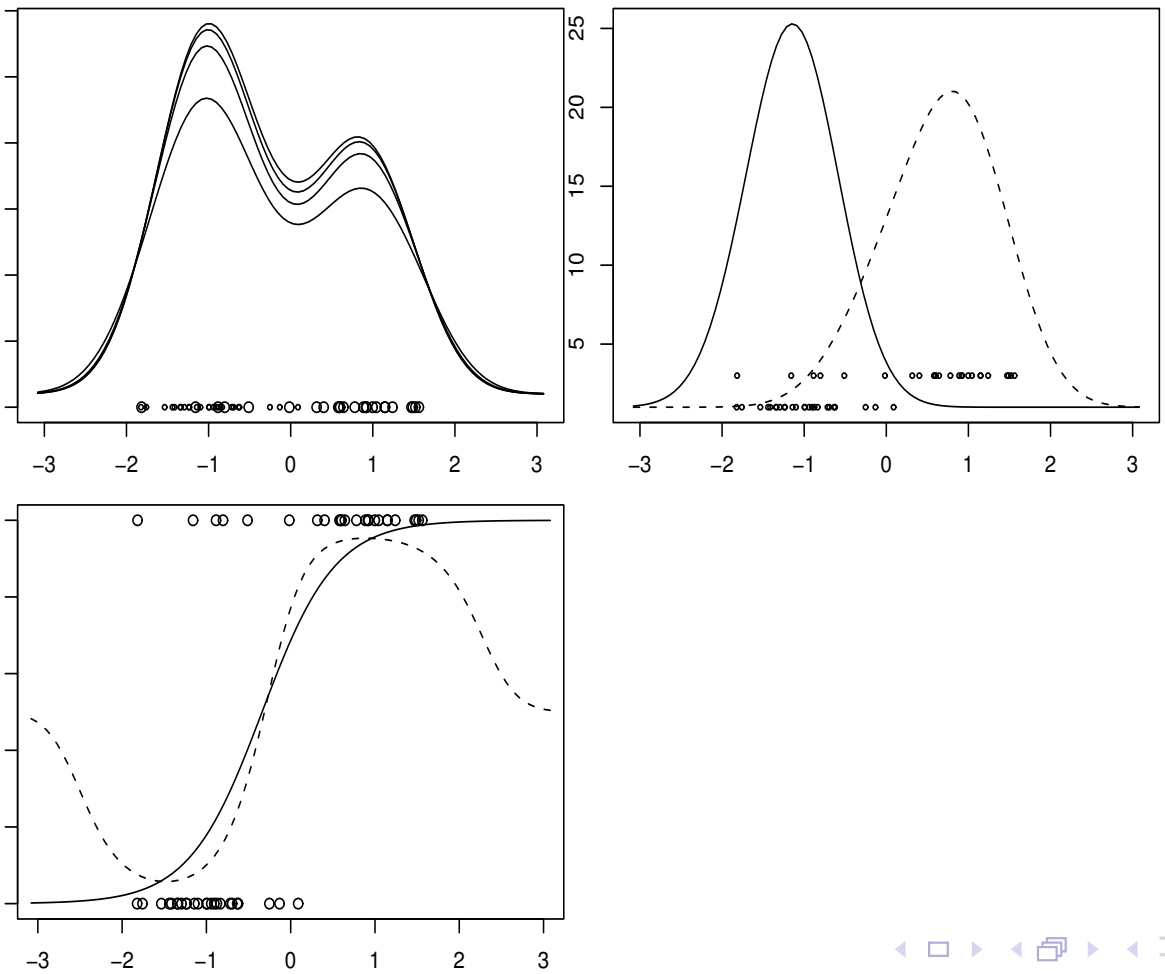
$$p_{n+1}(y(x') = r \mid \text{data}) \propto \mathrm{per}_{\alpha_r}[K](\mathbf{x}^{(r)}, x') / \mathrm{per}_{\alpha_r}[K](\mathbf{x}^{(r)})$$

Conditional relative intensity $\mathrm{per}_\alpha[K](\mathbf{x}, x')/\,\mathrm{per}_\alpha[K](\mathbf{x})$



$K(x, x') = \exp(-(x - x')^2)$ on $(0, 2\pi)$. Solid curve is for $\alpha = 1$, dashed curve for $\alpha = 0.25$, and the dotted curve for $\alpha = 0$.

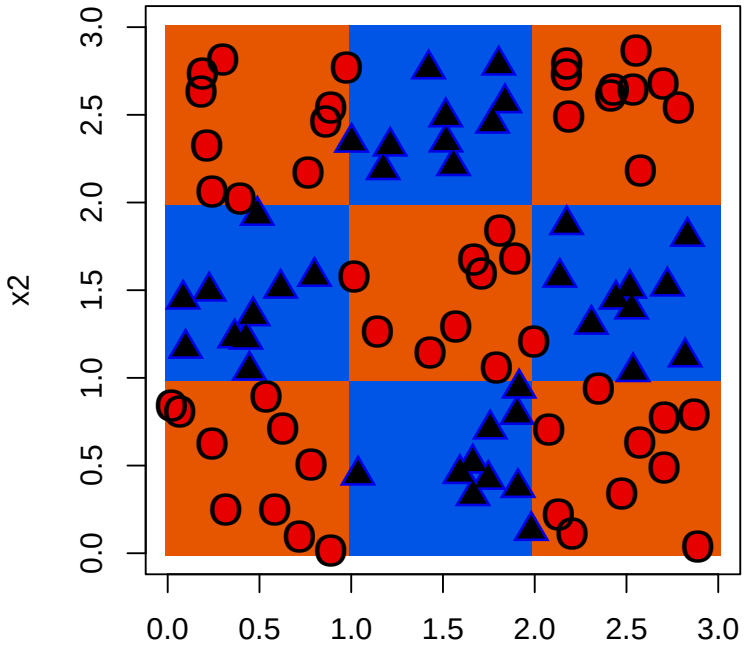
Illustration 2: Relative intensity functions $\mathrm{per}[K](\mathbf{x}.x')/\,\mathrm{per}[K](\mathbf{x})$ One-dimensional feature space: 2 labelled classes:



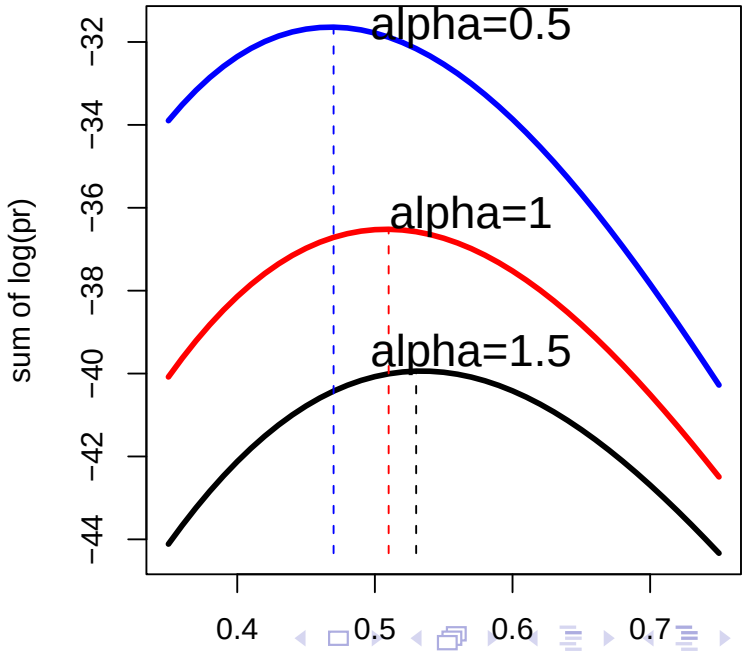
Chequerboard Pattern

A two-class example is generated by a 3×3 chequerboard pattern. The permanent model with $\alpha_1 = \alpha_2 = \alpha$ and $K(x, x') = \exp(-\|x - x'\|^2/\tau^2)$ is used.

Data Set



Cross Validation



Chequerboard Pattern: Classification